D358–D365 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 24 November 2012*
doi:10.1093/nar/gks1116

OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs

Robert M. Waterhouse[1,2], Fredrik Tegenfeldt[1,2], Jia Li[1,2], Evgeny M. Zdobnov[1,2,3] and Evgenia V. Kriventseva[1,2,*]

[1]Department of Genetic Medicine and Development, University of Geneva Medical School, [2]Swiss Institute of Bioinformatics, rue Michel-Servet 1, 1211 Geneva, Switzerland and [3]Division of Molecular Biosciences, Imperial College London, South Kensington Campus, London SW7 2AZ, UK

Received September 22, 2012; Revised October 19, 2012; Accepted October 21, 2012

ABSTRACT

The concept of orthology provides a foundation for formulating hypotheses on gene and genome evolution, and thus forms the cornerstone of comparative genomics, phylogenomics and metagenomics. We present the update of OrthoDB—the hierarchical catalog of orthologs (http://www.orthodb.org). From its conception, OrthoDB promoted delineation of orthologs at varying resolution by explicitly referring to the hierarchy of species radiations, now also adopted by other resources. The current release provides comprehensive coverage of animals and fungi representing 252 eukaryotic species, and is now extended to prokaryotes with the inclusion of 1115 bacteria. Functional annotations of orthologous groups are provided through mapping to InterPro, GO, OMIM and model organism phenotypes, with cross-references to major resources including UniProt, NCBI and FlyBase. Uniquely, OrthoDB provides computed evolutionary traits of orthologs, such as gene duplicability and loss profiles, divergence rates, sibling groups, and now extended with exon–intron architectures, syntenic orthologs and parent–child trees. The interactive web interface allows navigation along the species phylogenies, complex queries with various identifiers, annotation keywords and phrases, as well as with gene copy-number profiles and sequence homology searches. With the explosive growth of available data, OrthoDB also provides mapping of newly sequenced genomes and transcriptomes to the current orthologous groups.

INTRODUCTION

Homology in molecular biology refers to a common ancestry. In practice, homologous genes are recognized through the assessment of the statistical significance of sequence similarities of aligned nucleotides or amino acids. With reference to a specific species radiation, homologous relations define orthologs—'equivalent' genes in different species descended from a single ancestral gene (1–3). Speciation events, gene duplications, losses and sequence mutations lead to the diversity of genes encoded in the genomes of modern species. For any given set of species, all the descendants of a single gene from their last common ancestor constitute an orthologous group of genes. Orthology is therefore inherently hierarchical, referring explicitly to the last common ancestor, such that mostly one-to-one orthologs are identified among closely related species, whereas among more distantly related species orthologous groups comprise all surviving descendants of the ancestral gene.

There are two main approaches for orthology delineation: (i) algorithms that cluster all-against-all pairwise sequence comparisons, usually first identifying best-reciprocal matches between genomes that correspond to the shortest path over the speciation node of a distance-based tree, e.g. (4–12); and (ii) phylogeny-based methods that first define homologous gene families, build gene trees for each family, and then explicitly or implicitly reconcile them with the species tree often employing assumptions on rates of gene losses and duplications, e.g. (13–18). Phylogeny-based approaches have more parameters and may therefore yield better accuracy given sufficient data, but are often limited by the quality of multiple sequence alignments. This approach also considerably increases computational demands and becomes impractical for hundreds of species.

Recent benchmarking of prominent orthology resources (19,20) show that in the trade-off between specificity and sensitivity, OrthoDB assignments favor greater specificity with reasonable sensitivity, a balance that is well-suited to the goal of inferring gene functions. Although orthology is strictly an evolutionary concept, it can support the tentative transfer of functional annotations from well-studied organisms to orthologs in newly sequenced

*To whom correspondence should be addressed. Tel: +41 22 379 54 32; Fax: +41 22 379 57 06; Email: evgenia.kriventseva@isb-sib.ch

species. The confidence of such hypotheses on gene function may be qualitatively gauged by the genes' evolutionary histories, e.g. more confident inferences may be made for orthologs that are preserved across many species mostly as single-copy genes, with relatively low levels of sequence divergence, and consistent protein domain architectures. Gene duplicates in multi-copy orthologous groups often exhibit greater sequence divergence than single-copy orthologs (21), and as this may reflect biological innovation, any inferences on gene function should be made cautiously. OrthoDB classifications have proved to be accurate and biologically relevant as assessed within the framework of several recent genome projects, e.g. (22–26). Thus, the evolutionary characterization of orthologous groups in OrthoDB, collated with available gene functional annotations, provide a strong basis for making informed hypotheses that can drive evolutionary and molecular biology research.

SPECIES SAMPLING

The current OrthoDB release includes more than 250 eukaryotes and now also extends to cover prokaryotes with a total of 1115 bacterial species (Table 1, Supplementary Table S1). The predicted protein-coding gene sets and their corresponding General Feature Format (GFF) annotations for 52 vertebrate species were retrieved from Ensembl (27) (Release 67, May 2012). Data for the 45 arthropods were sourced from AphidBase (28), BeetleBase (29), FlyBase (30), Hymenoptera Genome Database (31), SilkDB (32), VectorBase (33), wFleaBase (34) and several genome consortia (as of July 2012). Gene sets for an additional 13 basal animal species were retrieved from Ensembl Genomes (35) and the Joint Genome Institute (36) (as of July 2012). The 142 fungal gene sets were retrieved from UniProt (37) (July 2012 release) and the bacteria were retrieved from NCBI (38) (Supplementary Table S1).

HIERARCHICAL ORTHOLOGOUS GROUPS

The OrthoDB orthology delineation procedure is based on clustering of best-reciprocal-hits (BRHs) between genes from each species pair, determined from all-against-all Smith–Waterman protein sequence comparisons now using SWIPE (39). The clustering procedure considers only the longest transcript per gene, and only the longest of all gene copies in a single genome with over 97% amino acid identity as determined by CD-HIT (40). Clusters are built progressively, with an e-value cutoff of 1e-3 for triangulating BRHs, and 1e-6 for pair-only BRHs, requiring an overall minimum sequence alignment overlap of 30 amino acids. The clusters of BRHs are subsequently further expanded to include all in-paralogs recognized as within-species homologs that are more closely related than the clustered BRHs.

Since its conception, OrthoDB (41) has promoted the concept of hierarchical orthology classifications by applying the clustering procedure at each radiation point of the considered species phylogeny and allowing users to explicitly select the most relevant level. It is rewarding to note that other resources e.g. (7,8) have embraced this concept and now provide orthology classifications at several major radiations across the tree of life. To determine the OrthoDB hierarchy, the species phylogenies in the current release were empirically computed using a maximum-likelihood approach as implemented in FastTree (42) over the super-alignment of mostly single-copy orthologs defined at the root node, multiply-aligned

Table 1. OrthoDB species and gene content

Lineage *Representative species*	Input genes		Classified genes (%)	Percentage of classified genes	
	Total	Average		in groups with annotation(s)[a]	in groups with phenotype(s)[b]
52 Vertebrates	951 245	18 293	92.7	96.3	48.4
Homo sapiens	20 827	na	94.9	93.5	45.6
Mus musculus	23 075	na	87.0	96.5	47.9
Danio rerio	26 206	na	80.7	96.9	48.5
45 Arthropods	746 324	16 585	71.1	87.1	25.1
Drosophila melanogaster	13 927	na	96.1	86.5	26.6
110[c] Metazoa	1 974 947	17 954	81.9	93.5	60.8
Caenorhabditis elegans	20 517	na	71.5	84.7	61.4
142 Fungi	1 223 848	8619	85.0	86.8	49.3
Saccharomyces cerevisiae	6652	na	96.2	91.9	94.8
1115 Bacteria	3 532 434	3168	91.0	91.6	47.1
Escherichia coli	4149	na	97.8	97.7	98.8
Haemophilus influenza	1657	na	98.2	98.8	85.3
Mycobacterium tuberculosis	3977	na	95.5	93.3	35.9

Statistics describing OrthoDB species coverage of vertebrate, arthropod, basal metazoan, fungal and bacterial orthologs with rich functional annotations.
[a]GO terms or InterPro domains.
[b]From Online Mendelian Inheritance in Man, the Mouse Genome Database, the Zebrafish Model Organism Database, FlyBase, WormBase, *Saccharomyces* Genome Database, EcoGene or the Database of Essential Genes.
[c]13 basal metazoan species plus 52 vertebrates and 45 arthropods.

using MAFFT (43), and filtered using TrimAl (44), and corroborated with known taxonomies from the literature.

The hierarchical orthology delineation procedure of the sampled lineages of vertebrates, arthropods and fungi classified 84% of a total of 2 921 417 protein-coding genes into 25 371, 33 393 and 55 793 orthologous groups, respectively (Table 1). Root-level delineation across the 110 animal species defined 58 308 orthologous groups covering 82% of the 3 198 795 metazoan genes and clustering of the 1115 bacteria classified 91% of the 3 532 434 bacterial genes. In addition to the root-level orthologs, 11 subgroups of bacteria—corresponding to the NCBI taxonomy 'class' levels—were clustered to provide more fine-grained orthologous groups for Actinobacteria, Spirochetes, Tenericutes, Thermotage, two classes of Cyanobacteria and Firmicutes, and three classes of Proteobacteria.

MAPPED FUNCTIONAL ANNOTATIONS

As orthologous groups comprise genes descended from a common ancestor, functional attributes ascribed to one or more members can be tentatively extrapolated to the last common ancestor and describe the group as a whole. In this way, orthologous group summary annotations provide an overview of mapped functional attributes with links to respective source databases to allow further investigations of the putative biological roles of their member genes (Figure 1).

Concise descriptors

Gene functional descriptions sourced from UniProt (37) and NCBI (38) provide succinct indications of known or inferred biological functions with coherent nomenclatures based on data from the literature as well as biocurator-evaluated and automatic computational classifications and annotations. In this OrthoDB release, frequently occurring phrases from member-gene descriptions label the group with a meaningful descriptor for each orthologous group.

Gene ontologies and InterPro domains

Molecular function, biological process and cellular component Gene Ontology (GO) (45) terms were retrieved from UniProt (37) and InterPro (46) protein domain signatures were sourced from the UniProt Archive of sequences. The available functional evidence for each orthologous group is summarized by listing the frequencies of associated GO terms and InterPro domains with concise attribute descriptions. Additionally, InterPro matches are displayed with domains ordered sequentially from the N- to C-terminus, describing the complete domain architecture of multi-domain genes, thereby allowing database queries with specific domain combinations. More than 85% of orthologs from each of the lineages are classified in groups that can be described by either GO terms or InterPro domains (Table 1).

Model organism phenotypes

OrthoDB gene annotations are enhanced with detailed functional data from well-studied model organisms in each lineage to highlight phenotypes associated with genes from *Mus musculus*, *Drosophila melanogaster* and *Saccharomyces cerevisiae*, sourced from the Mouse Genome Database (47), FlyBase (30) and *Saccharomyces* Genome Database (48), respectively. Eukaryotic model organism phenotypes now also include *Danio rerio* from the Zebrafish Model Organism Database (49) and *Caenorhabditis elegans* from WormBase (50). For bacteria, gene annotations are extended with phenotype data from EcoGene (51) for *Escherichia coli* genes and from the Database of Essential Genes (52) which covers 16 bacteria including *E. coli*, *Haemophilus influenza* and *Mycobacterium tuberculosis* (Table 1).

Online Mendelian inheritance in man

Human gene annotations are now enhanced with links to online Mendelian inheritance in man (OMIM®) (53), the catalog of associations between causative genes and human disease phenotypes, which describes thousands of allelic variants linked to numerous different disorders or susceptibilities. Mapping of human genes in OrthoDB to OMIM® records highlights known disease associations for almost 3000 genes (Table 1).

COMPUTED EVOLUTIONARY ANNOTATIONS

OrthoDB presents quantified orthologous group characteristics that describe evolutionary properties such as gene duplications or losses and rates of sequence divergence, these detail their evolutionary histories and provide a basis for the assessment of the confidence with which inferences on gene function may be made (Figure 1).

Phyletic profiles

Orthologous group phyletic profiles contrast the number of species with single-copy versus multi-copy orthologs and indicate the species coverage at the selected radiation point. The profiles thus highlight how descendant genes have been preserved across the phylogeny and whether gene duplications are widespread ('multi-copy license') or restricted ('single-copy control') as discussed in (21).

Evolutionary rates

The relative divergence among orthologous group member genes is quantified as the average of inter-species protein sequence identities normalized to the average identity of all inter-species BRHs. Appreciably higher or lower rates of divergence distinguish groups of orthologs with restrained or relaxed rates of protein sequence evolution, e.g. essential-gene-containing groups usually exhibit greater sequence conservation than those without.

Sibling groups

Homologous relations among genes from different orthologous groups at a given species radiation identify homologous or 'sibling' orthologous groups.

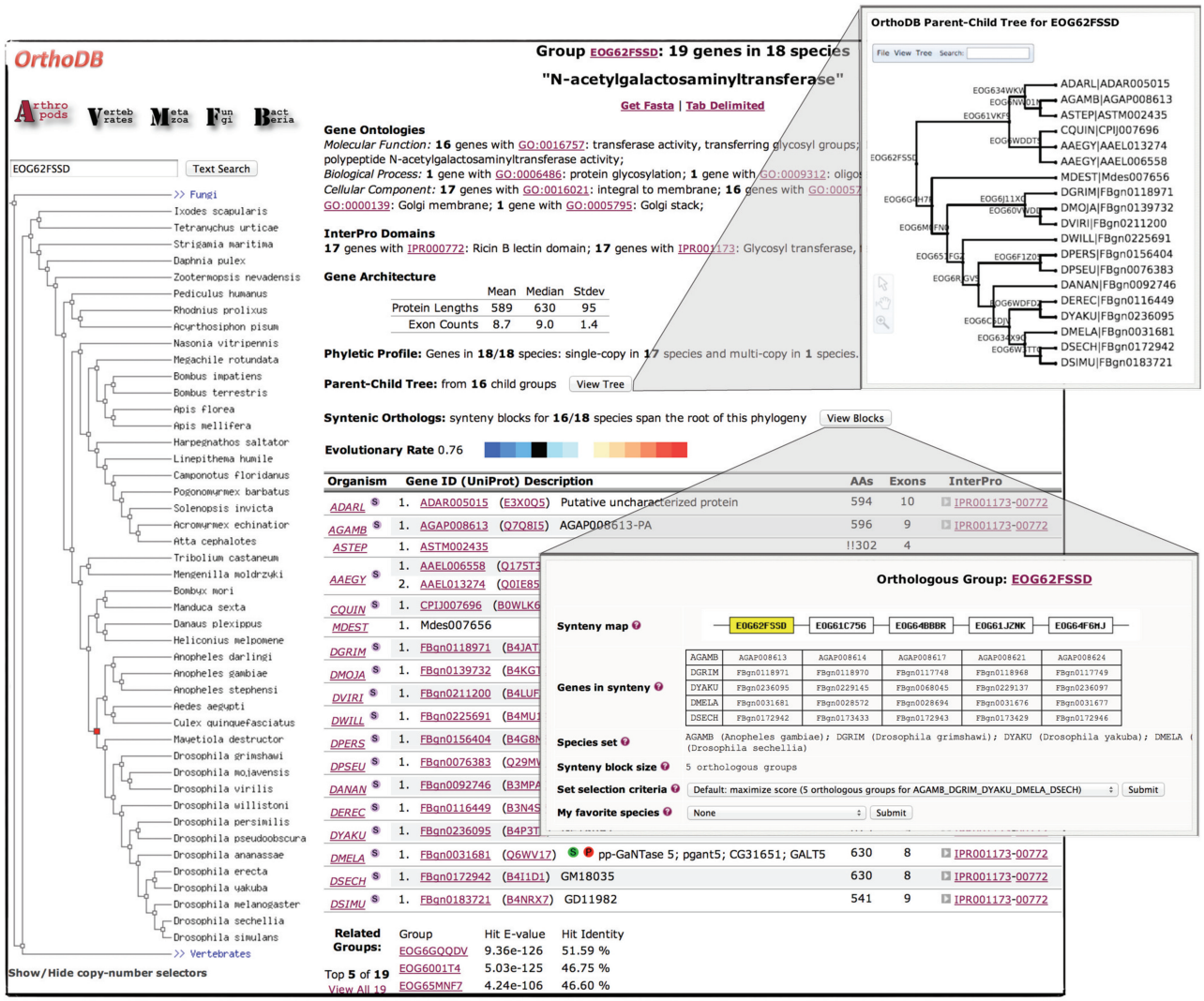

Figure 1. Screenshot of a sample orthologous group results page, featuring functional and evolutionary annotations, the inferred parent–child gene tree and syntenic orthologs.

These relations are quantified using data from all-against-all sequence comparisons by averaging over all pairs of homologs that link two orthologous groups with an e-value cutoff of 1e-3. This allows the user to retrieve sets of sibling orthologous groups that share significant sequence homology—which may therefore have some functional similarities—in an unbiased way that does not rely on protein domain or gene functional annotations.

Parent–child trees

Orthology delineation at each radiation along a given phylogeny hierarchically defines groups of orthologs with increasing resolution from the root level with the complete set of species to the most closely related species pairs. Parent–child relationships among orthologous groups delineated at each descendant radiation may therefore be defined by stepping along the phylogeny to identify orthologous groups with common subsets of genes (Figure 2). This new feature of OrthoDB represents these relationships as parent–child trees that illustrate the hierarchy of orthologous groups and their member genes, thereby building an inferred gene tree for a parent group by taking advantage of the greater resolution of its child groups. Users may view and edit the parent–child trees, as well as retrieve tree data formatted using Newick Utilities (54), from the 'Display Tree' window (Figure 1) that integrates the PhyloWidget (55) tool for the visualization and manipulation of phylogenetic tree data.

Gene architectures

Evolutionary annotations now also feature summary tables of protein lengths (all lineages) and exon counts (meatazoan lineages) that detail quantified mean, median and standard deviation values for each orthologous group, effectively describing a 'consensus' gene architecture. Amino acid and exon counts are also listed for each member gene, flagging those that are significantly shorter or longer than the consensus as potentially inaccurate gene model predictions.

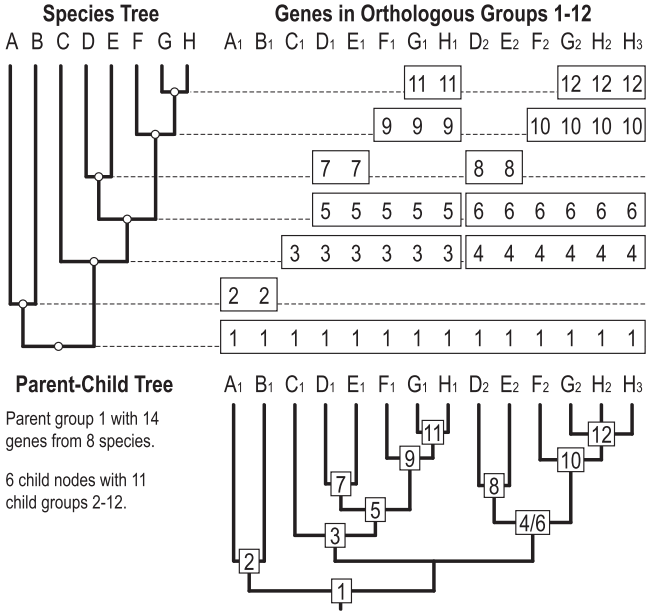

Figure 2. Hierarchical parent–child trees.

Syntenic orthologs

Comparing the chromosomal arrangements of orthologous genes among sets of species from the OrthoDB arthropod lineage identifies conserved blocks of syntenic orthologs. Such genes have maintained their local gene neighborhoods in the face of continual genomic evolution through sequence deletions, insertions and inversions, which may suggest selective advantages associated with their genomic arrangements, e.g. the TipE gene cluster of insect Para sodium channel auxiliary subunits (56). Ortholog-anchored synteny delineation (57) first identifies pairwise blocks with a minimum of two orthologs, allowing at most two intervening orthologs for each pair of genomes, and then successively projects these blocks through each pair of species across the phylogeny. The 'OrthoBlock' viewer (Figure 1) displays the best block—weighted according to the evolutionary span of the species and the number of orthologous groups in the block—selected from all the resulting blocks with at least five species for each orthologous group.

ORTHODB ONLINE

Selecting any species radiation point of interest from the interactive species trees, users can navigate through the hierarchy of orthologous groups defined at each radiation of the eukaryotic species phylogenies and for 11 major bacterial clades. At each orthology level, text searches return results from matches to various database identifiers and annotation keywords or phrases that can be combined through logical operator syntax to build more complex queries (e.g. ['cytochrome c'-mitochondrial]) using Sphinx indexing technology (http://sphinxsearch.com/). In addition, database cross-referencing of gene identifiers enhances search term matches through available gene names and synonyms, InterPro, or GO identifiers, as well as secondary identifiers from UniProt, Entrez GeneID, RefSeq, Protein Data Bank, OMIM, PubMed and model organism databases. Copy-number profile searches retrieve groups matching specific user-defined or general pre-defined phyletic profiles by combining the criteria of absent, present, single-copy, multi-copy or no restriction, for each species within any selected clade. BLAST (58) sequence similarity searches identify the best matches to genes from different species classified in OrthoDB, thereby allowing database querying with protein sequence data from any species. Importantly, although such sequence similarity searches with a single gene can recognize its homologs, accurate mapping to the defined orthologous groups requires assessment of the organism's complete gene set (see ortholog mapping section below). Searches stored during each user's web browser session provide a query history facility to allow recently executed queries to be reviewed, re-run or combined, e.g. a profile search for 'single-copy in >90% of species' could be combined with a text search with the GO identifier for 'receptor activity' to retrieve groups of mostly single-copy receptors. All search results may be easily exported as either Fasta-formatted files of protein sequences or tab-delimited text files of gene annotations, and the complete datasets are provided for download. All OrthoDB features are described in a comprehensive online help page and users may contact support@orthodb.org for additional information or specific requests, they may also subscribe to the low-traffic 'orthodb-news' mailing list (https://list.unige.ch/mailman/listinfo/orthodb-news) to keep abreast of the latest developments.

OrthoDB links

Search results present annotations for each orthologous group and tabulate all member genes with links to their respective sources e.g. Ensembl, UniProt, NCBI and FlyBase. Concise descriptors displayed for GO terms and InterPro domains are hyperlinked to their source records, and hyperlinks to OMIM and model organism databases provide direct access to all supporting data for genes with mapped phenotypes and synonyms. OrthoDB now provides FlyBase with orthology calls for the 12 *Drosophila* species as well as to selected arthropods and other animals. In addition, classified genes in OrthoDB are referenced with link-outs from UniProt records and NCBI gene link-outs.

Mapping of new species

Through a recently developed ortholog mapping procedure and corresponding web interfaces, OrthoDB now provides orthology classifications for genes from species with newly sequenced genomes mapped to existing orthologous groups. The mapping procedure first compares all genes from the new organism to all genes in OrthoDB groups, and then performs the BRH clustering procedure only allowing new genes to be added to existing clusters. The web interfaces list mapped genes and mirror OrthoDB data from the lineage(s) to which the new species is mapped. Thus, OrthoDB now provides online browsing

of mapped orthologs for new species with publically available gene sets such as the Chinese softshell turtle, *Pelodiscus sinensis*, (from Ensembl Release 68) (Supplementary Figure S1). Portals with restricted access provide the same functionality for private gene sets from organisms with recently sequenced genomes. For example, mapping the initial gene annotations of the genome of the alfalfa leafcutting bee, *Megachile rotundata*, helped to assess their quality and completeness, as well as providing a user-friendly portal to identify orthologs from other insects (G. Robinson, personal communication).

BENCHMARKING SETS OF UNIVERSAL SINGLE-COPY ORTHOLOGS

The fast-growing number of sequenced genomes and transcriptomes vary substantially in their completeness of sequencing, quality of read assembly and accuracy of gene annotation. A complementary approach to technical statistics such as the widely used N50 measure of genome assemblies, is to gauge the quality by examining the coverage of an expected gene set. This approach can assess not only completeness of genome coverage and fragmentation of the assembly, but also misassembly of haplotypes when the marker genes are known to exist only in single-copy, as well as the accuracy of annotation of such genes. For this purpose—of quality assessment of genomic data—we compiled benchmarking sets of universal single-copy orthologs (abbreviated BUSCOs) identified using OrthoDB for the Metazoan, Vertebrate, Arthropod and Fungal lineages (respectively, named BUSCO-Me, -Ve, -Ar, -Fu). Although these sets are intentionally conservative, they comprehensively sample each lineage and select representative genes from orthologous groups with single-copy orthologs in at least 90% of the species. The BUSCOs are available for download as Fasta-formatted protein sequences with corresponding gene, species and orthologous group identifiers.

PERSPECTIVES

The current OrthoDB release demonstrates the scalability of our computational procedures for the *ab initio* analysis of several millions of genes within a reasonable timeframe, e.g. with a 150 CPU-core computer cluster the total all-against-all sequence comparisons took about 1 month and the subsequent clustering procedures required from 1 day for the arthropod set to 4 weeks for the largest bacteria dataset on a single machine using a multi-threaded algorithm. Nevertheless, its comprehensive application to all emerging data will become prohibitive in a few years due to the exponential scaling of genome sequencing as well as to the variable completeness and quality of new genome annotations. Thus, our approach will be to focus the complete clustering analyses on only a representative selection of the best annotated species and those that maximize phylogenetic coverage, corroborating the results with curated classifications. These will form a comprehensive set of well-annotated and trusted orthologies to which genes from the other genomes, e.g. the thousands of insects to be sequenced through the i5K initiative (59), and new transcriptomes, e.g. from the 1KITE project (http://www.1kite.org), can be mapped.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1 and Supplementary Figure 1.

ACKNOWLEDGEMENTS

The authors would like to thank all members of the Computational Evolutionary Genomics Group and Dr Ivo Pedruzzi for useful discussions and suggestions, the Swiss Institute of Bioinformatics for pledging funds to support the maintenance and future development of OrthoDB, and the anonymous reviewers for their valuable comments and suggestions.

FUNDING

Swiss National Science Foundation [31003A-125350]; 'Commission Informatique' of the University of Geneva; and Schmidheiny Foundation. Funding for open access charge: Swiss Institute of Bioinformatics.

Conflict of interest statement. None declared.

REFERENCES

1. Fitch,W. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.
2. Koonin,E. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, **39**, 309–338.
3. Sonnhammer,E. and Koonin,E. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.*, **18**, 619–620.
4. Tatusov,R., Fedorova,N., Jackson,J., Jacobs,A., Kiryutin,B., Koonin,E., Krylov,D., Mazumder,R., Mekhedov,S., Nikolskaya,A. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinform.*, **4**, 41.
5. Chen,F., Mackey,A., Stoeckert,C.J. and Roos,D. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, **34**, D363–D368.
6. DeLuca,T.F., Cui,J., Jung,J.Y., St Gabriel,K.C. and Wall,D.P. (2012) Roundup 2.0: enabling comparative genomics for over 1800 genomes. *Bioinformatics*, **28**, 715–716.
7. Altenhoff,A.M., Schneider,A., Gonnet,G.H. and Dessimoz,C. (2011) OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res.*, **39**, D289–D294.
8. Powell,S., Szklarczyk,D., Trachana,K., Roth,A., Kuhn,M., Muller,J., Arnold,R., Rattei,T., Letunic,I., Doerks,T. *et al.* (2012) eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.*, **40**, D284–D289.
9. Ostlund,G., Schmitt,T., Forslund,K., Köstler,T., Messina,D., Roopra,S., Frings,O. and Sonnhammer,E. (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.*, **38**, D196–D203.
10. Yu,C., Desai,V., Cheng,L. and Reifman,J. (2012) QuartetS-DB: a large-scale orthology database for prokaryotes and eukaryotes inferred by evolutionary evidence. *BMC Bioinform.*, **13**, 143.
11. Waterhouse,R.M., Zdobnov,E.M., Tegenfeldt,F., Li,J. and Kriventseva,E.V. (2011) OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011. *Nucleic Acids Res.*, **39**, D283–D288.

12. Linard,B., Thompson,J.D., Poch,O. and Lecompte,O. (2011) OrthoInspector: comprehensive orthology analysis and visual exploration. *BMC Bioinform.*, **12**, Article 11.
13. Penel,S., Arigon,A.M., Dufayard,J.F., Sertier,A.S., Daubin,V., Duret,L., Gouy,M. and Perrière,G. (2009) Databases of homologous gene families for comparative genomics. *BMC Bioinform.*, **10(Suppl. 6)**, S3.
14. Huerta-Cepas,J., Capella-Gutierrez,S., Pryszcz,L.P., Denisov,I., Kormes,D., Marcet-Houben,M. and Gabaldón,T. (2011) PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res.*, **39**, D556–D560.
15. Ruan,J., Li,H., Chen,Z., Coghlan,A., Coin,L., Guo,Y., Hériché,J., Hu,Y., Kristiansen,K., Li,R. *et al.* (2008) TreeFam: 2008 Update. *Nucleic Acids Res.*, **36**, D735–D740.
16. Datta,R., Meacham,C., Samad,B., Neyer,C. and Sjölander,K. (2009) Berkeley PHOG: PhyloFacts orthology group prediction web server. *Nucleic Acids Res.*, **37**, W84–W89.
17. Vilella,A., Severin,J., Ureta-Vidal,A., Heng,L., Durbin,R. and Birney,E. (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**, 327–335.
18. Mi,H., Dong,Q., Muruganujan,A., Gaudet,P., Lewis,S. and Thomas,P. (2010) PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res.*, **38**, D204–D210.
19. Trachana,K., Larsson,T.A., Powell,S., Chen,W.-H., Doerks,T., Muller,J. and Bork,P. (2011) Orthology prediction methods: a quality assessment using curated protein families. *Bioessays*, **33**, 769–780.
20. Boeckmann,B., Robinson-Rechavi,M., Xenarios,I. and Dessimoz,C. (2011) Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees. *Brief. Bioinform.*, **12**, 423–435.
21. Waterhouse,R.M., Zdobnov,E.M. and Kriventseva,E.V. (2011) Correlating traits of gene retention, sequence divergence, duplicability and essentiality in vertebrates, arthropods, and fungi. *Genome Biol. Evol.*, **3**, 75–86.
22. Colbourne,J.K., Pfrender,M.E., Gilbert,D., Thomas,W.K., Tucker,A., Oakley,T.H., Tokishita,S., Aerts,A., Arnold,G.J., Basu,M.K. *et al.* (2011) The ecoresponsive genome of Daphnia pulex. *Science*, **331**, 555–561.
23. Werren,J.H., Richards,S., Desjardins,C.A., Niehuis,O., Gadau,J., Colbourne,J.K., Beukeboom,L.W., Desplan,C., Elsik,C.G., Grimmelikhuijzen,C.J.P. *et al.* (2010) Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. *Science*, **327**, 343–348.
24. Kirkness,E.F., Haas,B.J., Sun,W., Braig,H.R., Perotti,M.A., Clark,J.M., Lee,S.H., Robertson,H.M., Kennedy,R.C., Elhaik,E. *et al.* (2010) Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proc. Natl Acad. Sci. USA*, **107**, 12168–12173.
25. Arensburger,P., Megy,K., Waterhouse,R.M., Abrudan,J., Amedeo,P., Antelo,B., Bartholomay,L., Bidwell,S., Caler,E., Camara,F. *et al.* (2010) Sequencing of Culex quinquefasciatus establishes a platform for mosquito comparative genomics. *Science*, **330**, 86–88.
26. Bartholomay,L.C., Waterhouse,R.M., Mayhew,G.F., Campbell,C.L., Michel,K., Zou,Z., Ramirez,J.L., Das,S., Alvarez,K., Arensburger,P. *et al.* (2010) Pathogenomics of Culex quinquefasciatus and meta-analysis of infection responses to diverse pathogens. *Science*, **330**, 88–90.
27. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2012) Ensembl 2012. *Nucleic Acids Res.*, **40**, D84–D90.
28. Legeai,F., Shigenobu,S., Gauthier,J., Colbourne,J., Rispe,C., Collin,O., Richards,S., Wilson,A., Murphy,T. and Tagu,D. (2010) AphidBase: a centralized bioinformatic resource for annotation of the pea aphid genome. *Insect Mol. Biol.*, **19(Suppl. 2)**, 5–12.
29. Kim,H., Murphy,T., Xia,J., Caragea,D., Park,Y., Beeman,R., Lorenzen,M., Butcher,S., Manak,J. and Brown,S. (2010) BeetleBase in 2010: revisions to provide comprehensive genomic information for Tribolium castaneum. *Nucleic Acids Res.*, **38**, D437–D442.
30. McQuilton,P., St Pierre,S.E., Thurmond,J. and Consortium,F. (2012) FlyBase 101–the basics of navigating FlyBase. *Nucleic Acids Res.*, **40**, D706–D714.
31. Munoz-Torres,M.C., Reese,J.T., Childers,C.P., Bennett,A.K., Sundaram,J.P., Childs,K.L., Anzola,J.M., Milshina,N. and Elsik,C.G. (2011) Hymenoptera Genome Database: integrated community resources for insect species of the order Hymenoptera. *Nucleic Acids Res.*, **39**, D658–D662.
32. Duan,J., Li,R., Cheng,D., Fan,W., Zha,X., Cheng,T., Wu,Y., Wang,J., Mita,K., Xiang,Z. *et al.* (2010) SilkDB v2.0: a platform for silkworm (Bombyx mori) genome biology. *Nucleic Acids Res.*, **38**, D453–D456.
33. Megy,K., Emrich,S.J., Lawson,D., Campbell,D., Dialynas,E., Hughes,D.S., Koscielny,G., Louis,C., Maccallum,R.M., Redmond,S.N. *et al.* (2012) VectorBase: improvements to a bioinformatics resource for invertebrate vector genomics. *Nucleic Acids Res.*, **40**, D729–D734.
34. Colbourne,J., Singan,V. and Gilbert,D. (2005) wFleaBase: the Daphnia genome database. *BMC Bioinform.*, **6**, 45.
35. Kersey,P.J., Staines,D.M., Lawson,D., Kulesha,E., Derwent,P., Humphrey,J.C., Hughes,D.S., Keenan,S., Kerhornou,A., Koscielny,G. *et al.* (2012) Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Res.*, **40**, D91–D97.
36. Grigoriev,I.V., Nordberg,H., Shabalov,I., Aerts,A., Cantor,M., Goodstein,D., Kuo,A., Minovitsky,S., Nikitin,R., Ohm,R.A. *et al.* (2012) The genome portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Res.*, **40**, D26–D32.
37. UniProt-Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
38. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Federhen,S. *et al.* (2012) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **40**, D13–D25.
39. Rognes,T. (2011) Faster Smith-Waterman database searches with inter-sequence SIMD parallelisation. *BMC Bioinform.*, **12**, 221.
40. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
41. Kriventseva,E., Rahman,N., Espinosa,O. and Zdobnov,E. (2008) OrthoDB: the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res.*, **36**, D271–D275.
42. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
43. Katoh,K. and Toh,H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.*, **9**, 286–298.
44. Capella-Gutiérrez,S., Silla-Martínez,J.M. and Gabaldón,T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.
45. GO-Consortium. (2012) The Gene Ontology: enhancements for 2011. *Nucleic Acids Res.*, **40**, D559–D564.
46. Hunter,S., Jones,P., Mitchell,A., Apweiler,R., Attwood,T.K., Bateman,A., Bernard,T., Binns,D., Bork,P., Burge,S. *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.
47. Eppig,J.T., Blake,J.A., Bult,C.J., Kadin,J.A., Richardson,J.E. and Group,M.G.D. (2012) The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. *Nucleic Acids Res.*, **40**, D881–D886.
48. Cherry,J.M., Hong,E.L., Amundsen,C., Balakrishnan,R., Binkley,G., Chan,E.T., Christie,K.R., Costanzo,M.C., Dwight,S.S., Engel,S.R. *et al.* (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.*, **40**, D700–D705.
49. Bradford,Y., Conlin,T., Dunn,N., Fashena,D., Frazer,K., Howe,D.G., Knight,J., Mani,P., Martin,R., Moxon,S.A. *et al.* (2011) ZFIN: enhancements and updates to the Zebrafish Model Organism Database. *Nucleic Acids Res.*, **39**, D822–D829.
50. Yook,K., Harris,T.W., Bieri,T., Cabunoc,A., Chan,J., Chen,W.J., Davis,P., de la Cruz,N., Duong,A., Fang,R. *et al.* (2012)

WormBase 2012: more genomes, more data, new website. *Nucleic Acids Res.*, **40**, D735–D741.
51. Rudd,K.E. (2000) EcoGene: a genome sequence database for Escherichia coli K-12. *Nucleic Acids Res.*, **28**, 60–64.
52. Zhang,R. and Lin,Y. (2009) DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Res.*, **37**, D455–D458.
53. Amberger,J., Bocchini,C. and Hamosh,A. (2011) A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). *Hum. Mutat.*, **32**, 564–567.
54. Junier,T. and Zdobnov,E.M. (2010) The Newick utilities: high-throughput phylogenetic tree processing in the Unix shell. *Bioinformatics*, **26**, 1669–1670.
55. Jordan,G.E. and Piel,W.H. (2008) PhyloWidget: web-based visualizations for the tree of life. *Bioinformatics*, **24**, 1641–1642.
56. Li,J., Waterhouse,R.M. and Zdobnov,E.M. (2011) A remarkably stable TipE gene cluster: evolution of insect Para sodium channel auxiliary subunits. *BMC Evol. Biol.*, **11**, 337.
57. Zdobnov,E.M. and Bork,P. (2007) Quantification of insect genome divergence. *Trends Genet.*, **23**, 16–20.
58. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
59. Robinson,G.E., Hackett,K.J., Purcell-Miramontes,M., Brown,S.J., Evans,J.D., Goldsmith,M.R., Lawson,D., Okamuro,J., Robertson,H.M. and Schneider,D.J. (2011) Creating a buzz about insect genomes. *Science*, **331**, 1386.